## Тестовое задание для вакансии Machine Learning Developer. Предсказание индустрий вакансии.

В файле [train.xml.gz](https://drive.google.com/file/d/0B_KR2w6DQeX5NVlrNHRNY2UwaUk/view?usp=sharing) находится описание 10000 вакансий. Вакансии представлены в формате xml, в котором разные теги - различные поля вакансии. Тег industries содержит список индустрий, к которым относится вакансия. Одна вакансия может соответствовать нескольким индустриям. В файле [test.xml.gz](https://drive.google.com/file/d/0B_KR2w6DQeX5U3lKMzdNdk8tYlU/view?usp=sharing) содержится информация о других 10000 вакансий, но уже без индустрий. Необходимо обучить модель, предсказывающую список индустрий по полям вакансии. Качество предсказания оценивается с помощью метрики [mean F-score](https://www.kaggle.com/wiki/MeanFScore). Формат файла с предсказаниями индустрий для тестовой выборки должен быть следующим:

<vacancies>

<vacancy id="106034">

<industries>

<industry>...</industry>

<industry>...</industry>

...

</industries>

</vacancy>

<vacancy id="173332">

<industries>

<industry>...</industry>

<industry>...</industry>

...

</industries>

</vacancy>

...

</vacancies>

Мы ждём от вас вас исходный код и файл с предсказаниями. При оценке вашего решения мы будем обращать гораздо больше внимания на качество вашего кода, правильную оценку работы моделей, на основе которой вы будете выбирать решение. Качество финального предсказания для этого задания менее важно, мы не ждём сложных моделей, требующих много ресурсов и времени для обучения. Простое решение с приемлемым результатом предпочтительнее. Наше baseline решение "в лоб" даёт mean F-score=0.75548 на тестовой выборке, нас устроит совсем небольшое улучшение этого результата.

Задание необходимо решать на языке python c помощью библиотеки scikit-learn. Другие библиотеки для машинного обучения нежелательны. Вы можете пользоваться любыми библиотеками для частей кода, не связанных с обучением и оценкой качества моделей (например для парсинга входного xml).